SYSTEM AND METHOD FOR TRANSMITTING VOICE INPUT FROM A REMOTE LOCATION OVER A WIRELESS DATA CHANNEL

INVENTORS
Gilad Odinak
Thomas R. McCann
Julien Rivarol Vergin

PRIORITY CLAIM

This application claims priority from U.S. Provisional Application Ser. No. 60/280,379, filed March 29, 2001, Attorney Docket No. WING-1-1003.

FIELD OF THE INVENTION

This invention relates generally to the field of wireless communication, digital signal processing, and to a system and method for transmitting data derived from a voice input at a remote, mobile location over a wireless data channel.

BACKGROUND OF THE INVENTION

Cellular telephone transmission has long been used in mobile communication. Traditionally, cellular telephone transmission has been used to facilitate conversations between remote individuals. More recently, these same systems have been modified to facilitate the communication of verbal instructions to remote computer systems using speech recognition programs. In these modified systems, the speaker's verbal instructions are converted to digital data instructions, which in turn are used by a computer system to carry out the desired operation. The ability to communicate verbal instructions "hands-free" carries obvious advantages, not the least of which include safety advantages, especially when the speaker is concurrently attempting to operate a vehicle.

The traditional implementation of this type of speech transmission and recognition occurs in one of two ways: (1) sending raw audio (i.e., the speaker's verbal instructions) from a receiver in the vehicle, such as a cellular telephone, to the remote computer system, where the verbal instructions are converted to data instructions; or (2) performing extensive automated speech recognition (ASR) in the receiver (e.g., recognizing words and phrases), and sending the converted digital data instructions to the remote computer system. Both existing implementations suffer from significant disadvantages. Raw audio sent across a cellular network suffers from signal degradation, in turn diminishing the integrity of the audio signal to be converted to digital data instructions and, hence, the accuracy of the ultimate instructions. While converting the audio signal to digital data at the vehicle addresses this problem, it requires expensive computing power in the vehicle, which is logistically and cost prohibitive.

Thus, there is a need for a mobile speech recognition system and method that addresses the disadvantages with the current implementations.

SUMMARY

The present invention provides a voice communication system and method for improving voice recognition processing at a server system that receives voice input from a remotely located user system. The user system includes a microphone configured to receive user voice input, a processor configured to perform front-end voice recognition processing of the received user voice input, and a communication component configured to send the front-end processed user voice input to a destination over a network. The server system includes a communication component configured to receive the sent front-end processed user voice input, and a processor configured to complete voice recognition processing of the sent front-end processed user voice input.

In accordance with further aspects of the invention, the communication component of the user system communicates wirelessly.

In accordance with still further aspects of the invention, the processor of the user system includes a sampling component configured to sample the received user voice input. The processor of the user system further includes at least one of a noise cancellation component, an echo-cancellation component or an end-pointing component.

In accordance with yet other aspects of the invention, the processor of the server includes a component configured to perform a function based on the completed voice recognition processing.

In accordance with still another aspect of the invention, the user system includes removable modules. The modules include a processing module that includes a sampling component configured to sample the received user voice input. Also, the processing module includes a noise cancellation component, an echo-cancellation component or an end-pointing component.

In accordance with still further aspects of the invention, the modules include at least one of a positioning module, a phone adapter module, or a wireless network communication module.

As will be readily appreciated from the foregoing summary, the invention provides an improved system and method for performing voice recognition processing where the processing is performed remote from the person inputting the voice.

Brief Description of the Drawings

The preferred and alternative embodiments of the present invention are described in detail below with reference to the following drawings.

FIGURE 1 is a diagram illustrating the general architecture of a mobile speech recognition system that operates in accordance with the present invention;

FIGURE 2 is a flow chart illustrating the operation of the present invention;

FIGURE 3 is a flow chart illustrating the operation of one embodiment of the front-end sound processing aspect of the present invention; and

FIGURES 4A-C are block diagrams of data and voice package streams generated by an alternate embodiment of the present invention; and

FIGURE 5 is a flow diagram.

Detailed Description of the Preferred Embodiment

The present invention provides a system and method for transmitting data derived from voice input at a remote, mobile location over a wireless data channel. By way of overview and with reference to FIGURE 1, the mobile speech recognition system 10 of the present invention includes a user system 20 in wireless data communication with a server system 40 via a wireless data channel 60. The invention performs front-end sound processing in user system 20—short of full speech recognition conversion—and sends the results of the sound processing to server 40 in a digital form via wireless data channel 60.

More specifically, FIGURE 1 illustrates the particular components of the preferred embodiment of system 10 for sending preprocessed voice over the wireless data channel. User system 20 includes a user interface 22 having a microphone 24 for capturing the user's voice, a display 26, speakers 28, user interface 30 for allowing the user to perform various interactive functions, and preferably a video capture device 32. The user system further includes a voice processor 34 for performing front-end sound processing of the user's voice, and a communications device 36, such as an antenna or other signal carrier, for transmitting and receiving wireless information. User system 20 is preferably implemented in a mobile environment, such as in a car or other vehicle.

In an alternate embodiment, user system 20 includes a processor with memory coupled to multiple controllers (not shown) for receiving data from various vehicle data buses and module receptacles (not shown). The module receptacles are configured to receive various add-on modules, such as modules for coupling other electronic devices, communication modules, processing modules and other modules, such as a phone adapter module, a CPU/memory module, a bluetooth, infrared or wireless network communication module, and a global positioning system (GPS) module. In operation, voice capture and front-end sound processing is accomplished in various modules to the user system. Such modules are replaceable, allowing a user to receive up-to-date sound processing software or other more advanced modules without having to replace any hardware components already installed in the vehicle.

User system 20 is in communication with server system 40. The server system includes a server 42 for housing user system information, as well as processing and responding to requests for information from the user system and information sources 44, which may be integral with or independent from server system 40. The information sources store information subject to requests from the user system. In the preferred embodiment, the server includes a processor, a memory, and a database (not shown). Server 42 may be in communication with information sources 44 via direct access (e.g., hard-wired or point-to-point connection) as well as over Internet 46. Server system 40 further includes a means for sending and receiving information to user system 20, discussed below.

Wireless data channel 60 facilitates communication of instructions and information between user system 20 and server system 40. In a preferred embodiment, the wireless data channel may include a satellite system 62 in combination with a satellite dish 64 along with or in the place of an access point 66, the latter as part of a cellular or other wireless transmission network. In operation, instructions are transmitted from user system 20 via transmission/reception device 34 to either the satellite system or access point, which in turn communicate the instructions to server 42, in the former case via satellite dish 64. Conversely, information may be communicated from the server to the user system along a reverse direction of the same route.

The system and method of the present invention is better understood with reference to FIGURE 2, which illustrates the operational steps of the mobile speech recognition system. At block 100, the system receives the user's voice (acoustical audio input) at user system 20. At block 102, the system performs front-end sound processing on the acoustical audio input using processor 34, including converting the audio input to representative digital data. This front-end sound processing is discussed in more detail below with reference to FIGURE 3. Once front-end sound processing is complete, at block 104 the system packages the digital data for wireless delivery. At block 106, the system transmits the packaged digital voice information from user system 20 to server system 40 via wireless data channel 60.

At block 108, server system 40 performs complete speech recognition processing on the digital voice information using server 42. As part of this process, the server attempts to match the digital voice information with corresponding digital data entries in the server database, which in turn are associated with a program instructional language. One of three results follow from this process: (1) a match is found, in which case the corresponding program instructions are executed at block 110; (2) no match is found, in which case no instructions are executed (or partial instructions are executed) at block 110; or (3) multiple matches are found. In the last case, the server employs statistical models to decide which of the several matched entries is most probably the desired entry. This statistical modeling may look at previously matched digital voice information and, alone or in context with associated digital voice information, may combine entries in combination to best form appropriate instructional language (e.g., "the chair is white" is more probable than "the chair his white."). The server may also employ grammar, which is a subset of rules associated with the program instructional language, which includes the structure for valid sentences at certain points in the dialog from which the digital voice information was derived. Once the most probable match is determined, the resulting program instructions are executed at block 110.

The front-end sound processing of the present invention is better understood with reference to FIGURE 3. Microphone 24 of user system 20 captures the user's voice as acoustical audio input. At block 200, this acoustical audio input is converted to an analog electrical signal. At block 202, the analog signal is sampled and digitized. In the preferred embodiment, there are thousands of samples per second. Each sample is assigned a numerical value, which is typically between 0-255 so that it fits in a byte. In an effort to improve the accuracy of the resulting sample, processor 34 performs noise-cancellation, echo-cancellation, and end-pointing routines. Noise-cancellation involves an examination of the amount of energy in the analog signal at various frequencies. Processor 34 detects the background noise accompanying the original signal and subtracts it from the sample. Similar to noise-cancellation, echo-cancellation involves an examination of the amount of energy in the analog signal at various frequencies. With echo-cancellation, the goal is to reduce or eliminate from the sampled signal the vocal system prompt or audio component used by the system to elicit the user's original audio input. By comparing sampled signal components to a data stream representing the system prompt, processor 34 is able to subtract the system prompt and thereby separate the data stream representing the system prompt from the data stream representing the user's audio input. Also, undesired radio signals and other signals that processor 34 knows aren't the user's speech are also removed using the above described technique.

After performing noise- and echo-cancellation, processor 34 looks at the remaining amount of energy at the various frequencies and determines whether it contains actual user audio input, or user speech, as opposed to silence or other non-relevant noise. This is accomplished by examining the energy or amplitude of the audio input and the rate of "zero-crossing." Zero-crossing is where the audio signal changes from positive to negative, or vice versa. When the energy and zero-crossings are at certain predetermined levels, there is a likelihood of occurrence of a speech event.

In an alternate embodiment, processor 34 determines the beginning and end of a spoken phrase. This is performed by the technique known as end-pointing (i.e. speech detection). The determined beginning and end of a spoken phrase constitutes preprocessed information that is sent to server 40.

At block 204, resulting samples are grouped into sets representing acoustic audio for specified time periods. For example, in the preferred embodiment, a group contains all samples in a ten millisecond time period. The groups may have some overlap with respect to samples taken in succession, i.e., group N contains samples from time T to T+10 milliseconds; group N+1 contains samples from time T+8 to T+18 milliseconds, etc. In this example, if 20,000 samples are included per second, each group contains 200 samples.

At block 206, processor 34 converts grouped sets from time domain to frequency domain. In the preferred embodiment, this is accomplished using Fast Fourier Transform: the output of the transformation is a set of numbers, each set representing the amount of audio energy in a frequency band. The width of the frequency bands may be linear or nonlinear. For example, a given set of 10 millisecond time periods will have the following frequency domain values: 150Hz-300Hz: 75; 350Hz-400Hz: 100; . . . 1000Hz-1200Hz: 125 . . . 2500Hz-3000Hz: 47. In this example, the low frequency bands are 50Hz each, while the highest frequency band is 500Hz.

FIGURES 4A-C and 5 illustrate an alternate embodiment of the present invention for metering the deliverable information from user system 20 to server 42. FIGURES 4A-C illustrate packetized data streams 300, 310, 320 that are outputted streams during different modes of operation of user system 20. FIGURE 4A illustrates packet stream 300, which includes packets 304, 306 that are filled with voice information. Packet stream 300 also includes an address ID 302 located at the beginning. In this mode of operation, processor 34 has determined that the entire packet stream 300 should be filled with voice information and therefore will populate the packets with only voice information. This mode of operation occurs, for example, when the user is speaking into microphone 24.

FIGURE 4B illustrates a second mode of operation for the delivery of information from user system 20 to server 42. In this mode of operation, some packets 312 are filled with voice information and other packets 314 are filled with data, such as GPS or other vehicle information. Processor 34 operates in the mode shown in FIGURE 4B in order to deliver user system information along with the sent voice information. Periodically, vehicle information must be sent in order to be processed by the server for various functions. Often times it is desirable that vehicle information be associated with the voice information transmitted at a given time. In one example, a user may verbally request instructions to the nearest gas station, which answer would necessitate knowledge of the vehicle's present location. The present invention may send the voice instruction concurrent with GPS information, ultimately providing server 42 with both pieces of information as part of a nearly simultaneous transmission.

FIGURE 4C illustrates a mode of operation that processor 34 executes when no voice is being entered by the user, therefore allowing the entire packet stream 320 to be populated by data as shown in packets 322 and 324.

FIGURE 5 illustrates an example decision process performed by processor 34 when deciding how to populate packets sent from user system 20 to server 42. First, at decision block 350, processor 34 determines if user system 20 is receiving speech from the user. If it is determined that no speech is being received, processor 34 formats the transmission to include only data (see block 352), as shown by example in FIGURE 4C. If processor 34 determines that user system 20 is receiving speech, the process continues to decision block 356, where the processor determines if any data needs to be sent with the speech. If no data needs to be sent with the speech as determined by processor 34, the processor populates the packets of the transmission with only voice information (see block 358), as shown by example in FIGURE 4A. If processor 34 determines that data needs to be sent with the speech, the processor populates some of the data packets with voice and other data packets with data (see block 360), as shown by example in FIGURE 4B. When data packets and voice packets are distributed in the transmission, processor 34 populates packets with voice in order to optimize speech recognition at server 42.

While the preferred embodiment of the invention has been illustrated and described, as noted above, many changes can be made without departing from the spirit and scope of the invention. For example, the order of some or all of the front-end sound processing steps may be altered, such as the order in which noise-cancellation and echo-cancellation routines occur. In another example, the sequence used to determine whether input includes speech or data or both may be altered. Accordingly, the scope of the invention is not limited by the disclosure of the preferred embodiment.